# Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition

Maycel-Isaac Faraj and Josef Bigun, *Fellow*, *IEEE*

**Abstract**—This paper presents the scheme and evaluation of a robust audio-visual digit-and-speaker-recognition system using lip motion and speech biometrics. Moreover, a *liveness* verification barrier based on a person's lip movement is added to the system to guard against advanced spoofing attempts such as replayed videos. The acoustic and visual features are integrated at the feature level and evaluated first by a Support Vector Machine for digit and speaker identification and, then, by a Gaussian Mixture Model for speaker verification. Based on $\approx 300$ different personal identities, this paper represents, to our knowledge, the first extensive study investigating the added value of lip motion features for speaker and speech-recognition applications. Digit recognition and person-identification and verification experiments are conducted on the publicly available XM2VTS database showing favorable results (speaker verification is 98 percent, speaker identification is 100 percent, and digit identification is 83 percent to 100 percent).

**Index Terms**—Speech recognition, speaker recognition, motion estimation, normal image flow, normal image velocity, lip reading, lip motion, GMM, SVM, biometrics.

## 1 Introduction

The performance of currently available interactive recognition systems for speech and speakers is not robust against changes of environmental conditions—background noise, type of microphone, and so on. Using visual features can improve the recognition rate in acoustically noisy environments that have background noise or cross talk among speakers. Another advantage of combining audio and video for interactive person recognition is its ability to prevent unknown user attacks using prerecorded facial images or speech data of known people. In recent years, techniques have been suggested to combine visual and audio features to help solve recognition problems [7], [12], [33]. The performance of multimodal systems using audio and visual information is known to be superior to those of the acoustic and visual subsystems [3], [7], [32], [1]. Visual dynamic lip features can be used in speech and speaker-recognition systems to provide complementary information [39], [37], [23], leading to improved speaker-recognition performance, as demonstrated by [32] and [17]. Williams [38] suggested a method to capture facial parts with markers attached to the face and DeCarlo and Metaxas [10] presented a model-based full-face tracking system, though its relevance to speech recognition has not been shown. The lip information is suggested based on the shape and intensity of the lip region [28], [11], [19], [6], [21] because changes in the mouth shape, including the lips and tongue, carry significant phoneme-discrimination information. Luettin and Thacker [25] suggested shape models to represent changed mouth shapes as feature vectors—for example, gray-level distribution profiles around the lip contours. Finally, whole-word models are built by Hidden Markov Models (HMMs) for visual speech [25]. This solution relies exclusively on the accurate detection of the lip contours, often a challenging task under varying illumination conditions and rotations of the face. Another disadvantage is the fluctuating computation time due to the iterative convergence process of the contour extraction. The motion in dynamic lip images can be modeled by moving-line patterns [13], [20], also known as *normal image velocity*.

Integration by fusion has been increasingly studied in multimodal recognition systems [1], [35]. Here, we use direct feature fusion to obtain the audio-visual observation vectors by concatenating the audio and visual features. The resulting vectors are called observation sequences and are then modeled with a Gaussian Mixture Model (GMM) and a Support Vector Machine (SVM) classifier for speech and speaker recognition, respectively. Recent studies report good performance with SVMs used as classifiers in recognition [36], [15], [8]. An SVM can provide a powerful discriminative classifier for finding models of the boundary between a speaker and impostors compared to traditional methods for speaker recognition such as GMMs [29] and artificial neural networks [14]. By exploring SVMs, we study the performance influence of the classification method on speaker and speech recognition.

In this paper, we extend previous work [13] by introducing a new feature-reduction method used for quantization. Furthermore, we study these lip features in three novel types of applications—text-prompted audio-visual speech recognition (digit recognition), speaker recognition, and liveness detection.[1] The paper starts by briefly describing our feature-extraction technique for the mouth region along with our feature-reduction method in Section 2, followed by the acoustic feature extraction method in Section 3. The database used and the experimental setup are described in

• *The authors are with the School of Information Science, Computer, and Electrical Engineering, Halmstad University, PO Box 823, SE-301 18 Halmstad, Sweden. E-mail: {maycel.faraj, josef.bigun}@ide.hh.se.*

*Manuscript received 15 Feb. 2007; revised 24 Apr. 2007; accepted 26 Apr. 2007; published online 7 May 2007.*
*Recommended for acceptance by R.C. Guido, L. Deng, and S. Makino.*
*For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TC-0060-0207.*

1. A biometric system can increase its barriers against spoofing attacks by seeking evidence for *liveness*, that is, by attempting to detect if the biometric signal is captured from a physically present person as opposed to a photograph or tape recorder.

Section 4. Section 5 describes the GMM and SVM classifiers used for speech and speaker recognition, and the experimental results are shown in Section 6. We conclude with a discussion of the results and remaining issues in Section 7.

## 2 Visual Features by Normal Image Velocity

*Optical flow* is the distribution of apparent velocities in the movement of brightness patterns in an image sequence. Horn and Schunck [18] and Lucas and Kanade [22] presented widely used methods for determining the motion by linear regression among a family of optical flow analysis techniques. Horn and Schunck's method is based on an iterative algorithm that is more difficult to implement on simple computational architectures. Lucas and Kanade's method has the advantage of only using two-frame processing. This method is computationally inexpensive, but can be unreliable when estimating motion in edge displacements in image sequences because of its texture assumptions. Bigun et al. [2] proposed a different motion estimation technique based on an eigenvalue analysis of the multidimensional structure tensor, allowing the minimization process of fitting a line or a plane to the spectrum to be carried out without the Fourier transform. Applied to optical flow estimation and known as the 3D structure-tensor method, the eigenvector belonging to the largest eigenvalue of the tensor is directed in the direction of the contour motion, if such a motion is present. However, this method can be excessive for applications that need only line motion (contour motion) features. Assuming that the local neighborhood in the image contains lines or edges (not points or "textures"), the computations can instead be carried out in 2D subspaces of the 3D spatiotemporal space. For lip motion in image sequences, this assumption is realistic [13]. Lines in the spatiotemporal image will generate planes with a normal that can be estimated by using complex arithmetic and convolution.

A line in the image plane translated with a certain velocity in the normal direction will generate a plane in the spatiotemporal image. The velocity component of the translation parallel to the line cannot be calculated; this is referred to as the *aperture problem*. The normal unit vector is denoted as $\mathbf{k} = (k_x, k_y, k_t)^T$ and the projection of the normal vector to the $x$-$y$ coordinate axes represents the direction vector of the line's motion. The normal $\mathbf{k}$ of the plane will then relate to the velocity vector $v\mathbf{a}$ as follows:

$$\mathbf{v} = v\mathbf{a} = -\frac{k_t}{k_x^2 + k_y^2}(k_x, k_y)^T = -\frac{1}{(\frac{k_x}{k_t})^2 + (\frac{k_y}{k_t})^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^T, \quad (1)$$

where $\mathbf{v}$ is the *normal image flow*. The normal velocity estimation problem becomes a problem of solving the tilts $(\tan\gamma_1 = \frac{k_x}{k_t})$ and $(\tan\gamma_2 = \frac{k_y}{k_t})$ of the motion plane in the $xt$ and $yt$ manifolds, which is obtained from the eigenvalue analysis of the 2D structure tensor [2]. Using complex numbers and smoothing, the angles of the eigenvectors are given effectively by

$$\tilde{u}_1 = \int\int \left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial t}\right)^2 dxdt, \quad (2)$$

$$\tilde{u}_2 = \int\int \left(\frac{\partial f}{\partial y} + i\frac{\partial f}{\partial t}\right)^2 dydt. \quad (3)$$

Here, $\tilde{u}_1$ is complex valued $(i = \sqrt{-1})$ such that its magnitude is the difference between the eigenvalues of the local structure tensor in the $xt$ manifold, whereas its argument is twice the angle of the most significant eigenvector approximating $2\gamma_1$. The interpretation of $\tilde{u}_2$ is analogous to that of $\tilde{u}_1$. The function $f$ represents the continuous local image, whose sampled version can be obtained from the observed image sequence. Thus, the arguments of $\tilde{u}_1$ and $\tilde{u}_2$ deliver the *Total Least Square* estimations of $\gamma_1$ and $\gamma_2$ in the local 2D manifolds $xt$ and $yt$, respectively, but in the double-angle representation [16], leading to the following estimated velocity components:

$$\frac{k_x}{k_t} = \tan\gamma_1 = \tan\left(\frac{1}{2}\arg(\tilde{u}_1)\right) \Rightarrow \tilde{v}_x = \frac{\tan\gamma_1}{\tan^2\gamma_1 + \tan^2\gamma_2}, \quad (4)$$

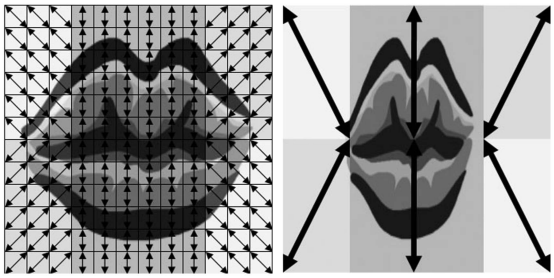
$$\frac{k_y}{k_t} = \tan\gamma_2 = \tan\left(\frac{1}{2}\arg(\tilde{u}_2)\right) \Rightarrow \tilde{v}_y = \frac{\tan\gamma_2}{\tan^2\gamma_1 + \tan^2\gamma_2}. \quad (5)$$

The tildes over $v_x$ and $v_y$ indicate that these quantities are estimations of $v_x$ and $v_y$. With the calculated 2D-velocity feature vectors $(v_x, v_y)^T$ in each mouth-region frame ($128 \times 128$ pixels), we have dense 2D-velocity vectors. Our earlier method [13] reduced the 2D-velocity vector to one dimension and then used a computationally exhaustive clustering method. Here, we suggest replacing the clustering method with a mean approximation, yielding a much faster and more intuitive algorithm for feature reduction. We reduce the 2D-velocity feature vectors $(v_x, v_y)^T$ at each pixel to one-dimensional scalars where the expected directions of motion are 0, 45, and $-45$ degrees—marked with three different gray-scale shades in six regions in Fig. 1a. The motion vectors within each region become real scalars that take the signs $+$ or $-$ depending on which direction they move relative to their expected spatial directions (differently shaded boxes):

$$f(p,q) = \left\|(v_x(p,q), v_y(p,q))\right\| * \mathrm{sgn}(\angle(v_x(p,q), v_y(p,q))). \quad (6)$$

Here, $p, q = 0 \ldots 127$.

Why use three spatial directions in six regions? This is because local lip motions are not completely free but must follow physical constraints. Mase and Pentland [26] and Yamamoto et al. [39] investigated lip articulation during speech by means of motion detection around different people's mouths. It is possible to conclude from these and other observations that the articulation of the lips progresses in a constrained manner during lip movement. For instance, when making the sound /o/, the lip articulators deform so that the right and left sides of the mouth move toward each other while the upper and lower areas move up and down, respectively (Fig. 1b).



(a) (b)

Fig. 1. (a) Illustration of velocity estimation quantification and reduction. (b) Lip articulation of the sound /o/.

The next step is to quantize the estimated velocities from arbitrary real scalars to a more limited set of values. Empirically, we found that direction and speed quantization significantly impacts the effect of noise on our estimate of the motion around the lip area. The quantized speeds are obtained from the data by calculating the mean value as follows:[2]

$$g(l,k) = \sum_{p,q=0}^{N-1} f(Nl+p, Nk+q). \quad (7)$$

Here, $p, q = 0 \ldots (N-1)$ and $l, k = 0 \ldots (M-1)$, where $N = 10$ and $M = 12$ represent the window size of the boxes (Fig. 1a) and the number of boxes, respectively. The resulting mean values were used as a feature vector representing 144-dimensional $(M \times M)$ feature vectors containing the statistics of lip motion. It is worth noting that the original dimension before reduction is $128 \times 128 \times 2 = 32,768$.

## 3 Acoustic Features

A person's vocal tract structure is a distinguishable physical property that is implicitly reflected in the speech spectrum. The Mel-Frequency Cepstral Coefficient (MFCC) is a commonly used instance of the filter-bank-based features [9] that can represent the speech spectrum. The speech features in this study were the MFCC vectors generated by the Hidden Markov Model Toolkit (HTK) [40] processing the data stream from the XM2VTS database. The use of an MFCC can be further motivated as the approximation of a basic psychophysical function in the human oral system reflected in the speech spectrum.

The input signal is preemphasized and divided into one 25-ms frame every 10 ms. A Hamming window is applied to each frame that is computed by MFCC vectors from the fast Fourier transform (FFT)-based Mel-warped log-amplitude filter bank, followed by a cosine transform and cepstral filtering. This MFCC vector contains 12 cepstral coefficients extracted from the Mel-frequency spectrum of the frame with normalized log energy, 13 delta coefficients (velocity), and 13 delta-delta coefficients (acceleration).

## 4 XM2VTS Database

All experiments in this paper use the XM2VTS database, currently the largest publicly available audio-visual database containing speech with faces [27]. The XM2VTS database contains image sequences and speech of 295 subjects (male and female), captured over four sessions. In each session, the subject is asked to pronounce three sentences when recording the video sequence; we use only "0 1 2 3 4 5 6 7 8 9." Two test protocols using GMM and SVM were applied in the experiments.

**Protocol 1.** This is the Lausanne protocol (Configuration 1) defined by the M2VTS consortium standardizing person-recognition experiments. It splits the database into training, evaluation, and test groups [24]. The evaluation set is used to quantify client and impostor access performance after training. The evaluation set is used to find the threshold for accepting or rejecting a person at predefined operation points. Finally, the test data is used to quantify how well the algorithm performs with respect to the desired performance once the thresholds are fixed. The training group contains 200 subjects as clients, the evaluation group contains an additional 25 subjects as impostors, and the testing group contains yet another 70 subjects as impostors. For clients, sessions 1, 2, and 3 are used as the training and evaluation sets and session 4 is used as the test set. This protocol is used in the person-verification (GMM) and person-identification (SVM) experiments below. For the XM2VTS database, the Lausanne protocol is commonly used as a standard protocol for speaker-identity experiments. However, no standard protocol is proposed for speech recognition by the M2VTS consortium. For our experiments on speech recognition, we use protocol 2, explained next. It is also used by other studies, such as that of Gavat et al. [15].

**Protocol 2.** It is worth noting that the XM2VTS data is difficult to use as is for speech recognition experiments because the speech and image sequences are not annotated. Therefore, before defining a protocol, we needed to annotate both the speech and visual data, which we did nearly 100 percent automatically by speech segmentation. For each person in the XM2VTS database, the "continuous" pronunciation "0 1 2 3 4 5 6 7 8 9" was divided into single digit subsequences of 0 to 9 using the HMM models. For our segmentation, we used HMM models of the digits. Furthermore, we manually verified and corrected the segmentation results so as to eliminate the impact of database segmentation errors when interpreting our recognition results. From the segmented database, we use 10 words (digits from 0 to 9) spoken by 295 speakers, each with eight pronunciations. The training and test group contains 295 subjects. Sessions 1 and 2 are used for the training group and sessions 3 and 4 are used for the test set. Although there was no specific identity modeling, the same speakers were used in both the training and the testing of the SVMs. The training samples we used were completely disjoint from the test samples. We used a total of four pronunciations for training and another four for testing (Fig. 2).

Through our semiautomatic segmentation, we noted that the words 4, 5, 8, and 9 contained much less visual data than the other digits during the speech utterance.

## 5 Classification

HMMs were introduced in speech recognition to provide a better model of the dynamic spectral features. When using an HMM in speaker recognition, the HMM is trained on a

2. Four pixels width boundary are removed $120 \times 120$.

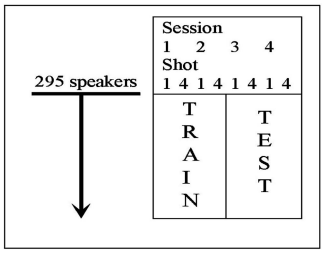



Fig. 2. Protocol 2 for digit identification.

chosen unit (phoneme, word, or sentence) by a target speaker (client). The query data are then compared to the modeled data according to verification (1:1 matching) or identification (1:many matching) tests. Depending on the topology of the HMM, a model can be either text dependent (sentences found in the training data set are also found in the test data set) or text independent (words or sentences found in training data do not necessarily reflect the words or sentences found in the testing data). A special-case HMM is the single-state (continuous-density) HMM; if the probability density function used for model observations in that state is a Gaussian mixture density, the model is usually called a GMM. Reynolds and Rose [30] introduced the GMM into the speaker-recognition field, reporting favorable results. A GMM is trained to optimize some criterion defined on the training data from a target speaker. This model then generates a likelihood for the query data when the system is operational. Another popular method is the SVM. SVMs were explored in speaker-recognition experiments with different types of kernels by Schmidt and Gish [31] and Wan and Campbell [36]. SVM is a discrimination-based binary classifier that normally models boundaries between two classes of training data in some (usually high-dimensional) feature space with no intermediate estimation of observation densities [34], [4]. An SVM is characterized mainly by its kernel function. Our observations indicate that SVM is a faster tool than GMM when the system is operational.

### 5.1 GMM

This probability model can be understood as a weighted sum of multivariate Gaussian distributions:

$$p(\mathbf{x}|\lambda) = \sum_{j=1} p_j b_j(\mathbf{x}). \tag{8}$$

Here, $\mathbf{x}$ is a D-dimensional feature vector and $p_j$ and $b_j(\mathbf{x})$ represent the mixture weights and the component densities, which are multivariate Gaussian densities. The weights $p_j$ represent the probability that identity $\lambda$, a person, is represented by the feature coming from a specific region of the feature space $\mathbf{x}$. In our system, we use the subword level (phonemes) using a GMM with five states and three mixtures in each state. Although it is reasonably simple to implement and it yields good performance, training and verification may take significant computation resources.

### 5.2 SVM

The SVM formulation is based on the Structural Risk Minimization principle, which minimizes an upper bound on the generalization error, as opposed to Empirical Risk Minimization [34], [4]. An SVM is a discrimination-based binary method using a statistical algorithm. It has good ability to generalize, which is why it has been used in pattern-recognition and information-retrieval tasks. The background idea in training an SVM system is finding a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ as a decision boundary between two classes. There exist techniques allowing the use of the fundamental SVM binary decision method in classification tasks with more than two classes.

For a linearly separable training data set of labeled pairs $\mathbf{x}_j, y_j,\ j = 1, \ldots, l$, where $\mathbf{x}_j \in \Re^n$ and $\mathbf{y} \in \{1, -1\}^l$, the following equation is verified for each observation (feature vector):

$$d_j(w^T \mathbf{x}_j + b) \geq 1 - \xi_j \quad \text{for } j = 1, 2, \ldots, l\ \xi_j > 0, \tag{9}$$

where $d_j$ is the label for sample $\mathbf{x}_j$, which can be $+1$ or $-1$, $\mathbf{w}_j$ and $b$ are the weights and bias that describe the hyperplane, and $\xi$ controls the number of data samples left inside the decision area, regulating the number of training errors. In our experiment, we use the inner product kernel function as the Radial Basis Function (RBF) kernel:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2), \gamma > 0. \tag{10}$$

When conducting speech-classification experiments, we will need to choose between multiple classes. The best method of extending the two-class classifiers to multiclass problems appears to be application dependent. *One against all* consists of building SVM classifiers equal to the number of classes. We train each SVM with one of the classes against the rest of the classes. The *one-against-one* approach simply constructs, for each pair of classes, an SVM classifier that separates those classes. All tests here were performed for only the speech signal, only the visual signal, and the merged audio-video signal using the SVM toolkit [5] *one-against-one* method. Based on the empirical evaluation of SVM with $\gamma = 2$ and $C = 1$, it gives optimal performance using only the training data.

## 6 Experiments

We want to quantify the performance of our visual features as a single or multimodality system in speaker recognition and speech recognition using the XM2VTS database. The experiments use the protocol setups from Section 4 with two classifiers. We use the feature-level direct-fusion technique. The feature integration technique can be found in more detail in the work of Faraj and Bigun [13]. First, the text-prompted speaker-verification system is presented, followed by the speaker-identification test using Protocol 1. Finally, we show the digit-identification system test results using Protocol 2.

### 6.1 Text-Prompted Speaker-Verification System by GMM

Table 1 presents the verification performance (with False Acceptance being equal to False Rejection in the evaluation set) of the acoustic, visual, and the combined audio-visual systems using Protocol 1. The verification performance was $\approx 77$ percent for the speaker-verification system based on only visual information. Speaker-verification based on the bimodal system gives $\approx 98$ percent correct verification, which is better than the single-modality system based on the audio or the visual information.

TABLE 1
Speaker-Verification Results of the Acoustic, Visual, and Merged Bimodal Audio-Visual System Using GMM

| Set / System | Evaluation | Test |
|---|---|---|
| Acoustic | 96% | 94% |
| Visual | 81% | 77% |
| Audio-Visual | 99% | 98% |

### 6.2 Speaker-Identification System by SVM

We perform experimental tests using just one word (digit) to recognize the speaker identity. The reason for using a single word is that an SVM has a tendency to become a computationally exhaustive machine for large data. However, it is possible to solve the volume problem by applying simple clustering, as in the work of Wan and Campbell [36], but that would introduce an unknown feature-selection method. The results of using SVM classifiers with an RBF kernel function to perform speaker recognition are shown in Table 2. For the experiment, we used a smaller data set of 100 speakers to see the most significant word for the speaker-recognition rate. Digit 7 gave the highest recognition rate, which we also tested for the whole database of 295 speakers (see Table 2). The experiment setup is given as follows:

- Partition the database for training, evaluation, and testing according to Protocol 1.
- Train the SVM for an utterance so that the classification score $L$ (the mean of the classification equation (9) for an utterance) is positive for the user and negative for impostors.
- $L$ is compared to a threshold $T$:
  - Find the threshold $T$ such that False Acceptance is equal to False Rejection using the evaluation set.
  - Using the threshold $T$, the decision $L$ is made according to the following rule: If $L > T$, accept the speaker; otherwise, reject her/him.

However, Protocol 1 is designed for verification (1:1 matching). To perform identification (1:many matching), we proceed as follows:

- Identify the speaker from a group of speakers:
  - We construct classifiers to separate each speaker from all other speakers in the training set.
  - The speaker identity is determined by the classifier that yields the largest likelihood score.

The performance obtained using bimodal recognition (100 percent) compared favorably with the classical single-modality recognition system based only on the speech signal (92 percent) and only on the visual signal (80 percent). When using coefficients from both the dynamic image and the speech, the recognition rates are considerably higher than when using a single modality based on speech parameters. These results indicate that our features can perform well in identification tasks.

TABLE 3
Speech-Identification Rate of All Digits Using Protocol 2 in a One against One SVM

| Word | Audio features | Visual features | Audio-Visual features |
|---|---|---|---|
| 0 | 89% | 70% | 92% |
| 1 | 90% | 77% | 100% |
| 2 | 86% | 60% | 89% |
| 3 | 90% | 75% | 96% |
| 4 | 89% | 55% | 85% |
| 5 | 90% | 50% | 83% |
| 6 | 100% | 90% | 100% |
| 7 | 93% | 100% | 100% |
| 8 | 91% | 54% | 83% |
| 9 | 90% | 49% | 85% |

### 6.3 Speech-Recognition System Using an SVM

We performed speech-recognition (digit-identification) tests according to Protocol 2. Using the bimodal system based on acoustic and visual-feature information, we obtain the best identification rate for digits 1, 6, and 7 (100 percent), as shown in Table 3. The results in Table 3 do vary, one cause being that there is not enough information (especially visual information) for certain utterances. This is not surprising because the XM2VTS database was collected for identity recognition and not speech recognition. The digits 4, 5, 8, 9, and, sometimes, 2 give worse identification rates in the bimodal mode than with just audio input for this reason. The lack of visual data has negatively influenced the bimodal fusion module, which presently assumes that the quality of information is uniform across the digits. During the segmentation, we could verify that, when uttering the words from 0 to 9 in a sequence without silence between words, the words 4, 5, 8, and 9 are pronounced in shorter time lapses and the visual data is notably less in comparison to other digits. Additionally, the duration of utterances for each speaker differs when uttering a word or digit. For instance, one speaker may take 10 image frames to utter one digit and another may take only four frames for the same digit. The average speech-recognition rate over all digits is $\approx 68$ percent and $\approx 90$ percent for visual and audio systems, respectively. The digit-identification rate using combined audio and video varies between 83 percent and 100 percent.

### 6.4 Liveness Detection

We illustrate in Table 4 the confusion matrix containing the digit-identification rate for one individual using only visual information using Protocol 1. The aim in this study is to interpret these probabilities of false/correct assignment as a *liveness* detection barrier. From the illustrated matrix in Table 4, we note that the probability estimation of correct assignment for deciding a digit 1 when it is uttered is 0.9.

Accordingly, the probability of deciding "non-i" when the signal is from an utterance of the digit "i" is thus called $\tilde{p}_i$, for instance, $\tilde{p}_1 = (1 - p_1)$ for the digit 1 is 0.1. *Liveness* is a relatively new research area in biometrics and there are

TABLE 2
Person Identification Rates by SVM Using the Word "7"

| Kernel | Audio identification rate | Visual identification rate | Audio-Visual identification rate |
|---|---|---|---|
| RBF | 92% | 80% | 100% |

TABLE 4
Confusion Matrix for the Digit Identification for One Person "370"

| - | 0 | 1 | 2 | 3 | 7 |
|---|---|---|---|---|---|
| 0 | 0.9 | 0 | 0 | 0.10 | 0 |
| 1 | 0 | 0.63 | 0 | 0.27 | 0 |
| 2 | 0 | 0 | 0.78 | 0.22 | 0 |
| 3 | 0 | 0 | 0 | 1.00 | 0 |
| 7 | 0 | 0 | 0 | 0.41 | 0.59 |

several strategies to decide on *liveness*. Our strategy is to use the recognized digits in the lip-reading (machine) expert discussed above on what has been uttered, as a *liveness* evidence or score. For instance, the *liveness* score for a machine expert recognizing four times of five digits is 0.8, representing the average identification rate. The *liveness* assessment expert for a random digit sequence will be

$$p = \sum_{i=0}^{I} (1 - \tilde{p}_i), \tag{11}$$

where $I$ is the total number of digits. The rationale behind this strategy is as follows: The *liveness* system knows the correct answer to be uttered by the clients because the digit is prompted by the system. Consequently, the system can measure the correct digit-identification rate of a random sequence from the utterances of the tested person. Effectively, if the identification rate of the machine is significantly less than the statistically expected $p$, the tested person may be judged as not live. In Table 4, the average identification of a digit based on (11) is approximately 0.78 using five digits, $\tilde{p}$ being approximately 0.22. It is also possible to add or remove the digits that are most difficult to recognize from the prompt sequence to increase the digit-identification chance of the machine, which in turn reduces the probability to accept a recorded sequence as live because the prerecorded sequence will not correspond to the random prompt sequence.

The *liveness* system can be extended by several experts, for instance, by a speech expert by summation fusion. For example, $w_i p_i + (1 - w_i) q_i$, where $p_i$ and $q_i$ are the visual expert and the speech expert and $w_i$ is a reliability measure in $[0, 1]$ to prefer the visual expert over the speech expert. The *liveness* score for the combined lip-motion and speech system would be

$$L = \frac{\sum_{i=0}^{I} (w_i p_i + (1 - w_i) q_i)}{I}. \tag{12}$$

The parameter $w_i$ can be chosen statistically based on how well the individual systems assess *liveness* independent of the person.

## 7 Discussion

We have described a system using lip motion in dynamic image sequences of various speakers for robust speech and speaker recognition without the use of iterative algorithms or assuming successful lip-contour tracking. Our results show that, through motion features, we can achieve improved results for speaker and speech recognition by fusing the audio and video signals at the feature level.

The experimental results confirm the importance of the visual signal as complementary information to speech. Furthermore, by utilizing only lip-motion information, we have presented a lip-*liveness* expert with experimental results on how to indicate *liveness* to raise a barrier against spoofing attempts such as replayed videos or photographs used as masks. In environments having heavy (acoustic) noise (for example, airports, trains, and airplanes), the recognition system based on only speech would, with high probability, fail. Accordingly, these experiments support the view of the added value of lip motion. A speaker or speech-recognition system deployed in a noisy environment (for example, an airport) could therefore potentially gain a reduction of its error rate if sufficient lip-motion information is added.

A speech-and-speaker-recognition system is presented, based on only visual features and also on the joint modalities of audio-visual features. Visual features from dynamic images differ considerably across speakers due to different looks and different mouth movements, which makes speech and speaker recognition very difficult from only images of lips. This is, to the best of our knowledge, the first study presenting lip-motion features as standalone and as a complement for speaker and speech recognition using extensive tests. We have provided reasonable experimental results that motion features contain rich information to achieve good recognition performance for speech, as well as identity, but, most importantly, in conjunction with audio features.

A significant bottleneck in lip-motion research became visible in our study, namely, the lack of a database that contains sufficient amounts of visual information on lip movements. The results indicate that, once the visual-feature extraction is made on a sufficient amount of visual speech data, the available modeling for recognition tasks is highly successful. Future work will therefore include suitable database construction with a larger vocabulary and with more visual information.

## Acknowledgments

The authors gratefully acknowledge the support of Halmstad University and the Swedish Research Council.

## References

[1] E. Bigun, J. Bigun, B. Duc, and S. Fischer, "Expert Conciliation for Multi Modal Person Authentication Systems by Bayesian Statistics," *Proc. First Int'l Conf. Audio- and Video-Based Person Authentication (AVBPA '97)*, J. Bigun, G. Chollet, and G. Borgefors, eds., pp. 291-300, 1997.

[2] J. Bigun, G. Granlund, and J. Wiklund, "Multidimensional Orientation Estimation with Applications to Texture Analysis of Optical Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 775-790, Aug. 1991.

[3] K.R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, Oct. 1995.

[4] C.J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM—A Library for Support Vector Machines," www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

[6] T. Chen, "Audiovisual Speech Processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9-21, 2001.

[7] C. Chibelushi, F. Deravi, and J. Mason, "A Review of Speech-Based Bimodal Recognition," *IEEE Trans. Multimedia,* vol. 4, no. 1, pp. 23-37, 2002.
[8] P. Clarkson and P. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '99),* vol. 2, pp. 585-588, 1999.
[9] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 357-366, 1980.
[10] D. DeCarlo and D. Metaxas, "Optical Flow Constraints on Deformable Models with Applications to Face Tracking," *Int'l J. Computer Vision,* vol. 38, no. 2, pp. 99-127, 2000.
[11] U. Dieckmann, P. Plankensteiner, and T. Wagner, "Acoustic-Labial Speaker Verification," *Proc. First Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA '97),* pp. 301-310, 1997.
[12] B. Duc, S. Fischer, and J. Bigun, "Face Authentication with Sparse Grid Gabor Information," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '97),* vol. 4, no. 21, pp. 3053-3056, 1997.
[13] M.I. Faraj and J. Bigun, "Person Verification by Lip-Motion," *Proc. Conf. Computer Vision and Pattern Recognition Workshop (CVPRW '06),* pp. 37-45, 2006.
[14] K. Farrell, R. Mammone, and K. Assaleh, "Speaker Recognition Using Neural Networks and Conventional Classifiers," *IEEE Trans. Speech and Audio Processing,* vol. 2, no. 1, pp. 194-205, 1994.
[15] I. Gavat, G. Costache, and C. Iancu, "Robust Speech Recognizer Using Multiclass SVM," *Proc. Seventh Seminar Neural Network Applications in Electrical Eng. (NEUREL '04),* pp. 63-66, 2004.
[16] G.H. Granlund, "In Search of a General Picture Processing Operator," *Computer Graphics and Image Processing,* vol. 8, no. 2, pp. 155-173, 1978.
[17] T.J. Hazen, "Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, no. 3, pp. 1082-1089, 2006.
[18] B. Horn and B. Schunck, "Determining Optical Flow," *J. Artificial Intelligence,* vol. 17, no. 1, pp. 185-203, 1981.
[19] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-Labial Speaker Verification," *Proc. First Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA '97),* pp. 319-326, 1997.
[20] K. Kollreider, H. Fronthaler, and J. Bigun, "Evaluating Liveness by Face Images and the Structure Tensor," *Proc. Fourth IEEE Workshop Automatic Identification Advanced Technologies (AutoID '05),* pp. 75-80, 2005.
[21] L. Liang, X.L.Y. Zhao, X. Pi, and A. Nefian, "Speaker Independent Audio-Visual Continuous Speech Recognition," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME '02),* vol. 2, pp. 26-29, 2002.
[22] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 674-679, 1981.
[23] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration Strategies for Audio-Visual Speech Processing: Applied to Text-Dependent Speaker Recognition," *IEEE Trans. Multimedia,* vol. 7, no. 3, pp. 495-506, 2005.
[24] J. Luettin and G. Maitre, "Evaluation Protocol for the Extended M2VTS Database *xm2vtsdb*," IDIAP Communication 98-054, Technical Report R R-21, number = IDIAP - 1998, 1998.
[25] J. Luettin and N. Thacker, "Speechreading Using Probabilistic Models," *Computer Vision and Image Understanding,* vol. 65, no. 2, pp. 163-178, 1997.
[26] K. Mase and A. Pentland, "Automatic Lip-Reading by Optical-Flow Analysis," *Systems and Computers in Japan,* vol. 22, no. 6, pp. 67-76, 1991.
[27] K. Messer, J. Matas, J. Kittler, and J. Luettin, "Xm2vtsdb: The Extended M2VTS Database," *Proc. Second Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA '99),* pp. 72-77, 1999.
[28] E. Petajan, B. Bischoff, D. Bodoff, and N.M. Brooke, "An Improved Automatic Lipreading System to Enhance Speech Recognition," *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI '88),* pp. 19-25, 1988.
[29] D. Reynolds, T. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, nos. 1-3, pp. 19-41, 2000.
[30] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models," *IEEE Trans. Speech and Audio Processing,* vol. 3, no. 1, pp. 72-83, 1995.
[31] M. Schmidt and H. Gish, "Speaker Identification via Support Vector Classifiers," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '96),* pp. 105-108, 1996.
[32] X. Tang and X. Li, "Fusion of Audio-Visual Information Integrated Speech Processing," *Proc. Third Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA '01),* pp. 127-143, 2001.
[33] X. Tang and X. Li, "Video Based Face Recognition Using Multiple Classifiers," *Proc. Sixth IEEE Int'l Conf. Automatic Face and Gesture Recognition (FGR '04),* pp. 345-349, 2004.
[34] V.N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995.
[35] P. Varshney, "Multisensor Data Fusion," *Electronics and Comm. Eng. J.,* vol. 9, no. 6, pp. 245-253, 1997.
[36] V. Wan and W. Campbell, "Support Vector Machines for Speaker Verification and Identification," *Proc. IEEE Signal Processing Soc. Workshop Neural Networks for Signal Processing X,* vol. 2, pp. 775-784, 2000.
[37] T. Wark, S. Sridharan, and V. Chandran, "The Use of Speech and Lip Modalities for Robust Speaker Verification under Adverse Conditions," *Proc. IEEE Int'l Conf. Multimedia Computing and Systems (ICMCS '99),* vol. 1, 1999.
[38] L. Williams, "Performance-Driven Facial Animation," *Proc. SIGGRAPH '90,* pp. 235-242, 1990.
[39] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip Movement Synthesis from Speech Based on Hidden Markov Models," *J. Speech Comm.,* vol. 26, no. 1, pp. 105-115, 1998.
[40] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0),* http://htk.eng.cam.ac.uk/docs/docs.shtml, 2000.

**Maycel Isaac Faraj** received the BSc and MSc (in 2003) degrees in computer science engineering from Halmstad University, Sweden, and the Licentiate of technology degree in signals and systems from Chalmers University of Technology, Sweden, in 2006. Since 2004, he has been a doctoral student in the Intelligent System Laboratory at Halmstad University. His research interests include biometrics, signal analysis, computer graphics, computer vision, and their applications in human-machine interaction.

**Josef Bigun** received the MSc and PhD degrees from Linköping University, Sweden, in 1983 and 1988, respectively. In 1988, he joined the Swiss Federal Institute of Technology in Lausanne (EPFL), where he worked until 1998 as an "adjoint scientifique." He was elected as a professor to the Signal Analysis Chair, his current position, at Halmstad University and the Chalmers Institute of Technology, Sweden, in 1998. His professional interests include biometrics, texture analysis, motion analysis, and the understanding of biological processing of audiovisual signals, including human face recognition. He is an elected fellow of the IAPR and the IEEE. He has contributed to the organization of several international conferences as a cochair or track chair, including the initiation of the Audio and Video-Based Biometric Person Authentication (AVBPA) conference series and the organization of several International Conferences on Pattern Recognition (ICPRs) and International Conferences on Image Processing (ICIPs). He has been an editorial member of several journals in the areas of pattern recognition and image understanding.